Customer Shopping Behaviour Analysis

1. Project Overview

The project analyses the business problem of understanding the customer shopping behaviour that gives an idea about the transactional information about the customers. The dataset includes 3900 rows about the transactional data. The goal is to understand the shopping patterns, repeated categories, category wise sales and revenue generated, also focusing on the age group factor for shopping.

2. Dataset Summary

- Rows: 3,900

- Columns: 18

- Key Features: - Customer demographics (Age, Gender, Location, Subscription Status)

- Purchase details (Item Purchased, Category, Purchase Amount, Season, Size, Color)

- Shopping behavior (Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases, Review Rating, Shipping Type)

- Missing Data: 37 values in Review Rating column

3. Exploratory Data Analysis using Python

- EDA was initialised with Data cleaning and preparation using Python (Jupyter notebook)
- Data loading using Pandas library.
- Used df.info() to gain some idea about the structure of the dataset and df.describe() to get a summary of the data present in the .csv file.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Customer ID             3900 non-null   int64
 1   Age                     3900 non-null   int64
 2   Gender                  3900 non-null   object
 3   Item Purchased          3900 non-null   object
 4   Category                3900 non-null   object
 5   Purchase Amount (USD)   3900 non-null   int64
 6   Location                3900 non-null   object
 7   Size                    3900 non-null   object
 8   Color                   3900 non-null   object
 9   Season                  3900 non-null   object
 10  Review Rating           3863 non-null   float64
 11  Subscription Status     3900 non-null   object
 12  Shipping Type           3900 non-null   object
 13  Discount Applied        3900 non-null   object
 14  Promo Code Used         3900 non-null   object
 15  Previous Purchases      3900 non-null   int64
 16  Payment Method          3900 non-null   object
 17  Frequency of Purchases  3900 non-null   object
dtypes: float64(1), int64(4), object(13)
memory usage: 548.6+ KB
```

```
df.describe(include='all')
```

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2	2
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No	No
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223	2223
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN	NaN

- Review rating column had a few missing data values. Those missing values were replaced by category-wise median values.
- Renamed all the columns to have better readability and accessibility.
- Created age-groups by categorizing the customer ages as mentioned in the .csv file. Added another column to get an idea of purchase frequency days for each consumer.
- As there was no difference in the columns discount_applied and promo_code_used, promo code column was dropped to remove redundancy as a part of data cleaning process.
- Connected Python script to MySQL and loaded the cleaned Data Frame into the database for SQL analysis.

```
from sqlalchemy import create_engine
from urllib.parse import quote_plus

# connecting to MySQL
username = [redacted]
password = quote_plus("[redacted]")
host = "[redacted]"
port = "[redacted]"
database = "vendor"

engine = create_engine(f"mysql+pymysql://{username}:{password}@{host}:{port}/{database}")

# Dataframe to MySQL table
table_name = "vendor_table"
df.to_sql(table_name, engine, if_exists = "replace", index = False)
```

4. Data Analysis using SQL (Solving the Business Problems)

- Gender-wise Revenue

gender	revenue
Male	157890
Female	75191

-Users using discount but spending more than average purchase amount

customer_id	location	purchase_amount
2	Maine	64
3	Massachusetts	73
4	Rhode Island	90
7	Montana	85
9	West Virginia	97
12	Hawaii	68
13	Delaware	72

-Top 5 highest average review ratings products

item_purchased	avg_review_rating
Gloves	3.86
Sandals	3.84
Boots	3.82
Hat	3.8
Skirt	3.78

-Comparison of shipping types and purchases through them

avg_purchase_amount	shipping_type
60.4752	Express
58.4602	Standard

-Subscribers v/s non-subscribers

subscription_status	customers	avg_spend	total_revenue
Yes	1053	59.4919	62645
No	2847	59.8651	170436

-Top 5 products with highest percentage of purchases with discounts applied

item_purchased	percentage_of_purchase
Hat	50.0000
Sneakers	49.6552
Coat	49.0683
Sweater	48.1707
Pants	47.3684

-New, Returning and Loyal customer bases

customer_segment	customers_count
New	83
Returning	701
Loyal	3116

-Top 3 most purchased products in each category

category	item_purchased	total_orders	ranking
Accessories	Jewelry	171	1
Accessories	Sunglasses	161	2
Accessories	Belt	161	3
Clothing	Blouse	171	1
Clothing	Pants	171	2
Clothing	Shirt	169	3
Footwear	Sandals	160	1
Footwear	Shoes	150	2
Footwear	Sneakers	145	3
Outerwear	Jacket	163	1
Outerwear	Coat	161	2

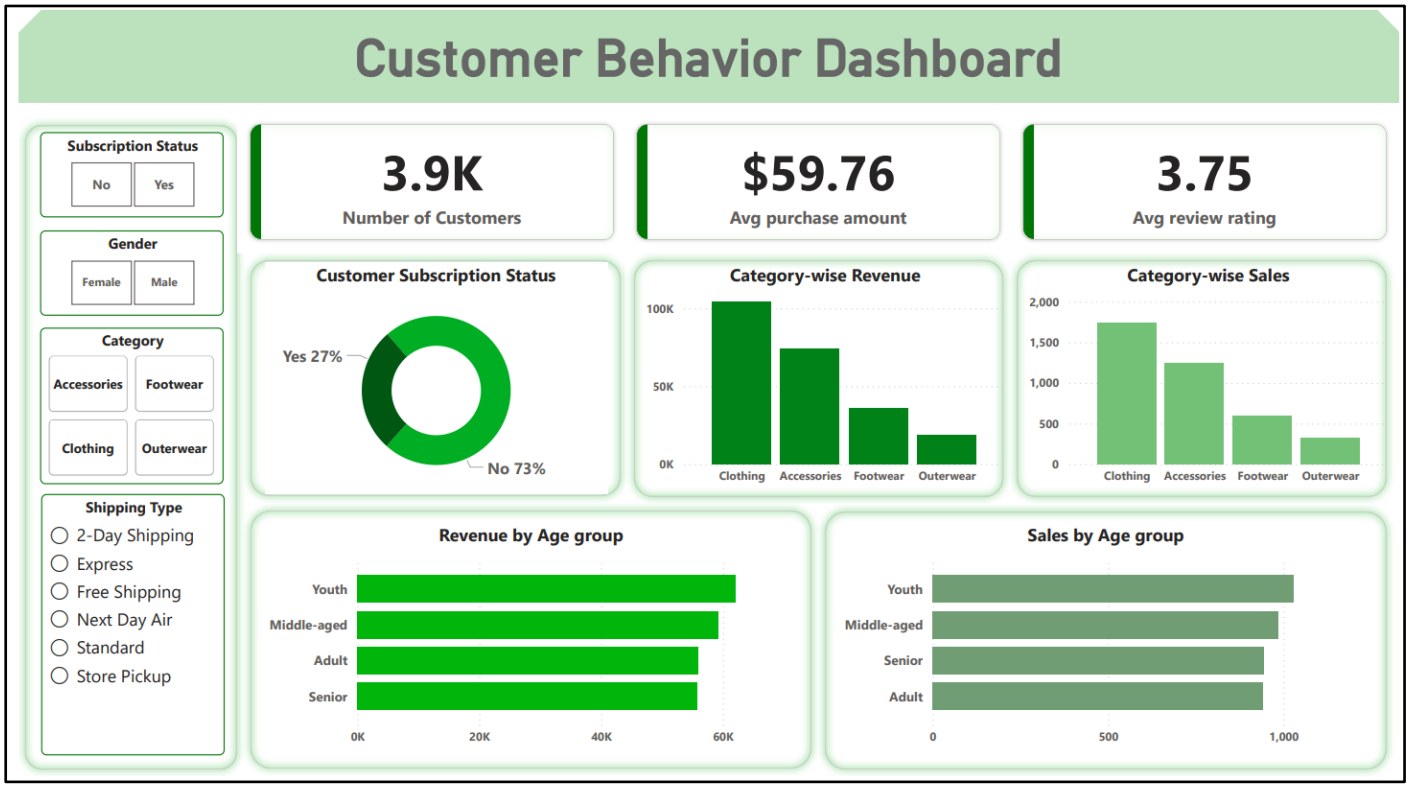
-Repeat buyers and subscribers relations

repeat_buyers	subscription_status
958	Yes
2518	No

-Age group wise Revenue

age_group	revenue
Youth	62143
Middle-aged	59197
Adult	55978
Senior	55763

5. Building Dashboard in Power BI



6. Derived Business Recommendations from the analysis of the data

- Boost Subscriptions – Promote exclusive benefits for subscribers.
- Loyal Customer Reward – Reward ‘Repeat’ buyers to move them into the ‘Loyal’ segment.
- Reframing Discount Strategy – Balance sales boosts with margin control.
- Product Positioning – Highlight top-rated and best-selling products in campaigns.
- Focused Marketing – Emphasized efforts on high-revenue age groups and express-shipping users.